For Every Generalization Action, Is There Really an Equal and Opposite Reaction? Analysis of the Conservation Law for Generalization Performance

R. Bharat Rao*
Electrical Engineering
University of Illinois
Urbana, IL 61801

Diana Gordon
Naval Research Laboratory
Washington, DC 20375-5337
gordon@aic.nrl.navy.mil

William Spears
Naval Research Laboratory
Washington, DC 20375-5337
spears@aic.nrl.navy.mil

Abstract

The "Conservation Law for Generalization Performance" [Schaffer, 1994] states that for any learning algorithm and bias, "generalization is a zero-sum enterprise." In this paper we study the law and show that while the law is true, the manner in which the Conservation Law adds up generalization performance over all target concepts, without regard to the probability with which each concept occurs, is relevant only in a uniformly random universe. We then introduce a more meaningful measure of generalization, *expected generalization performance*. Unlike the Conservation Law's measure of generalization performance (which is, in essence, defined to be zero), expected generalization performance is conserved only when certain symmetric properties hold in our universe. There is no reason to believe, a priori, that such symmetries exist; learning algorithms may well exhibit non-zero (expected) generalization performance.

1 INTRODUCTION

The theoretical analysis of inductive learning algorithms over all learning situations has been the subject of some recent research [Wolpert, 1992; Schaffer, 1993; Wolpert, 1994]. This paper begins by focusing on a recent result for concept learning, the "Conservation Law for Generalization Performance" [Schaffer, 1994]. This law states that for any learning algorithm and bias, "positive performance in some learning situations must be balanced by negative performance in others." The Conservation Law (henceforth, CLGP) has been likened by its author to other natural laws of conservation, and has attracted considerable attention in the learning community. In this paper, we study this law to understand its implications for inductive learning. (The CLGP is a reiteration of earlier "no-free-lunch" theorems developed by Wolpert [1992, 1994].)

In Section 2 we perform a rational reconstruction of the proof of the CLGP. This proof is implicitly the same as in Schaffer [1994], but makes explicit the fact that getting zero generalization performance depends on only one thing: the CLGP's uniform summation over target concepts (as in Wolpert [1994]). We later use this reconstruction to show that the way the CLGP sums generalization performance is relevant only in a uniformly random universe. This indicates that the CLGP, while trivially true, is not particularly relevant for inductive learning.

In a uniformly random universe, learning is necessarily impossible. We are interested in characterizing the properties of universes in which learning is indeed impossible. To this end, in Section 3 we present a more meaningful measure of generalization, *expected generalization performance* ($\mathcal{EGP}$), which measures the expected performance of a learner. This measure, unlike the interpretation of the CLGP in Schaffer [1994], does not preclude the existence of a general bias for learning in a universe. We then characterize the conditions under which *expected* generalization performance will be conserved: each of these require certain symmetries to exist. In Section 4 we propose a criterion for all learners for determining how *close* our universe comes to matching the required symmetry for zero $\mathcal{EGP}$.

2 CONSERVATION LAW REVISITED

In this section we show that the CLGP is equivalent to the statement:

> Labeling an unseen example as positive (or negative) results in a generalization accuracy of

[*] This paper will appear in the proceedings of the 1995 Machine Learning Conference. The current address for the first author is: Learning Systems Department, Siemens Corporate Research, 755 College Rd. East, Princeton, NJ 08540. E-mail: *bharat@scr.siemens.com*.

0.5, when the generalization accuracy is (effectively defined as being) measured *uniformly* against both possible classifications (i.e., positive or negative).

2.1 THE CONSERVATION LAW

We use notation similar to that used in the CLGP paper, namely, that there is a finite set of m possible attribute vectors or cases, and a target concept C that classifies each of these cases into one of two classes, C_0 and C_1.[1] Using a sampling distribution, $\mathcal{D}$, we draw n samples and classify them according to C to form a training set, θ.[2] A learner, $\mathcal{L}$, generalizes from θ to produce a concept that can classify previously unseen cases. Like the CLGP, we consider the learner's accuracy only on cases that are outside the training set, i.e., cases with attribute vectors not in θ. We distinguish between performance and accuracy where performance is identical to an analogous accuracy measure reduced by 0.5, i.e., performance is the improvement over random guessing.

A learning situation, S, is the triple $(\mathcal{D}, C, n)$. The *generalization accuracy* of a learner $\mathcal{L}$ in a learning situation S is denoted by $\mathcal{GA}(\mathcal{L}, S)$: it is defined as $\mathcal{L}$'s accuracy, with respect to C, over all unseen cases (weighted according to $\mathcal{D}$). (We provide a precise definition of generalization accuracy later in Equation 5.) The *generalization performance* of $\mathcal{L}$ is $\mathcal{GP}(\mathcal{L}, S) = \mathcal{GA}(\mathcal{L}, S) - 0.5$. Throughout this paper, the term "learner" (or learning algorithm) is equivalent to the *bias* [Mitchell, 1980; Utgoff, 1986] used by $\mathcal{L}$ to generalize from θ to unseen cases, and "$\forall\mathcal{L}$" should be read as "for all learning biases."

The CLGP states that in a classification problem, for any learning algorithm the total generalization performance over all learning situations is zero.

$$\forall\mathcal{L} \sum_{S}[\mathcal{GP}(\mathcal{L}, S)] = \forall\mathcal{L} \sum_{S}[\mathcal{GA}(\mathcal{L}, S) - 0.5] = 0$$

$$\forall\mathcal{L} \sum_{\mathcal{D}}\sum_{C}\sum_{n}[\mathcal{GA}(\mathcal{L}, (\mathcal{D}, C, n)) - 0.5] = 0 \quad (1)$$

Henceforth, we will drop the parentheses around $(\mathcal{D}, C, n)$ and use the notation, $\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, n)$. Later in this section, we introduce a new term, the *average generalization accuracy*, which sums the total generalization accuracy over all concepts. Upon restating the CLGP in terms of this new metric, we find that the CLGP sums up a large number of terms, each of which is zero (by the very definition of the CLGP). In the rest of this section, we will demonstrate that the CLGP (Equation 1) is equivalent to Equation 10, which we restate below,

$$\forall\mathcal{L} \sum_{\mathcal{D}}\sum_{n}\sum_{\theta}(prob(\theta) \cdot (0)) = 0$$

In short, by summing generalization over all target concepts, the CLGP virtually *defines* $\mathcal{GP}$ to be 0 for all θ, n, and $\mathcal{D}$, and the summation over training set sizes and sampling distributions (in Equation 1) is redundant.

2.2 ANALYZING THE CONSERVATION LAW

We begin by expanding the summation over training set sizes, and then focus on the crux of the CLGP: uniformly summing over all target concepts.

Summing over training set sizes, n: The generalization accuracy for an arbitrary n, $\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, n)$, is computed by averaging $\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta)$ over all θ's of size n. We can expand that term:

$$\sum_{n}[\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, n) - 0.5] \;=\; \sum_{n}\;\sum_{\theta\, of size\, n}(prob(\theta) \cdot [\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5])$$

where $prob(\theta)$ is the probability of drawing the n samples in the training set, θ, from the given sampling distribution, $\mathcal{D}$. By substituting in Equation 1, the CLGP can be rewritten as:

$$\forall\mathcal{L} \sum_{\mathcal{D}}\sum_{C}\sum_{n}\sum_{\theta}(prob(\theta) \cdot [\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5]) = 0$$

If a concept, C, classifies at least one of the samples in θ differently from the observed sample, then $prob(\theta) = 0$ for all such concepts, C. Interchanging the order of summation, we get:[3]

$$\forall\mathcal{L} \sum_{\mathcal{D}}\sum_{n}\sum_{\theta}(prob(\theta) \cdot \sum_{C}[\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5]) = 0 \quad (2)$$

Summing over all target concepts, C: The summation over all target concepts C in Equation 2 above is the *critical feature* of the CLGP. For an arbitrary $\mathcal{L}$, $\mathcal{D}$, n, and θ (of size n) this can be written as:

$$\sum_{C}[\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5] = \sum_{C}[\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta)] - 0.5 \cdot \sum_{C}\{1\}$$

$\sum_{C}\{1\} = |C|$ is the *total number of possible target concepts* that are consistent with the given θ. Note

[1] Continuous attributes can be discretized to any arbitrary degree of accuracy. The results in this paper can be trivially extended to cover multiple-class prediction as well.

[2] We do not assume $\mathcal{D}$ is *iid* or restrict it in any other way. Furthermore, $\mathcal{D}$ could depend on C: for example, in an active learning scenario, the classification of the cases already in θ can affect the probability with which future cases are drawn into θ.

[3] Strictly speaking, we should write $\sum_{C/\theta}$ in Equation 2, where C/θ is the set of all possible concepts which are consistent with the classification of samples in θ.

that the first term above ($\sum_C \mathcal{GA}\ldots$) is the total generalization accuracy; that is, the generalization accuracy of the concept learned by $\mathcal{L}$ (for the given θ and $\mathcal{D}$) summed over all possible targets. This can be rewritten as:

$$\sum_C \mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta) - 0.5\cdot|C| = \\ |C|\cdot\left(\frac{\sum_C \mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta)}{|C|} - 0.5\right)$$

The first term within the parentheses above ($\sum_C \mathcal{GA}(\ldots)/|C|$) is the total generalization accuracy over all targets divided by the number of targets, namely, the *average generalization accuracy* of a learner, $\mathcal{L}$. It is denoted by $\mathcal{AGA}_{\mathcal{L}}(\mathcal{D},\theta)$, or more simply $\mathcal{AGA}_{\mathcal{L}}$. To isolate the manner in which the CLGP sums generalization over all target concepts, we restate the CLGP in terms of $\mathcal{AGA}_{\mathcal{L}}$ (dividing both sides of the CLGP in Equation 2 by $|C|$):

$$\forall\mathcal{L}\quad \sum_{\mathcal{D}}\sum_{n}\sum_{\theta}(prob(\theta)[\mathcal{AGA}_{\mathcal{L}}(\mathcal{D},\theta) - 0.5]) = 0$$
$$\text{where}\quad \mathcal{AGA}_{\mathcal{L}}(\mathcal{D},\theta) = \frac{\sum_C \mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta)}{|C|} \qquad (3)$$

Computing the generalization accuracy: $\mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta)$ is defined as $\mathcal{L}$'s accuracy over all unseen cases with respect to C; this depends on how $\mathcal{L}$ maps unseen cases to classes. Consider an arbitrary training set, θ, of n cases which are labeled by a target, C, and are drawn from a sample distribution, $\mathcal{D}$. Let $\mathcal{Z}(\theta)$ (or more simply, $\mathcal{Z}$) be the set of all unseen cases, let $k(\theta)$ (or more simply, k) be the number of distinct cases in $\mathcal{Z}$, and $e_i \in \mathcal{Z}, 1 \leq i \leq k$.[4] We define a *label*, f, as a mapping $f : \mathcal{Z} \mapsto \{C_0, C_1\}$. (Throughout this paper we will use f or f_j to refer to arbitrary labels, and l to refer to labels learned by $\mathcal{L}$.)

For illustrative purposes we make three simplifying assumptions: that given θ, $\mathcal{L}$ learns a single label classifying the unseen cases, $l = \mathcal{L}(\theta)$, and that l and C are not stochastic. (We relax these assumptions in Appendix A.) As l maps each $e_i \in \mathcal{Z}$ to either C_0 or C_1, l can be represented by a k-bit binary vector:[5]

$$l = \langle b_1 b_2 \ldots b_k\rangle \quad b_i \; = \; \begin{cases} 1 & \text{if } l(e_i) = C_1 \\ 0 & \text{if } l(e_i) = C_0 \end{cases} \qquad (4)$$

For the particular target concept, C, used to label all cases, perfect generalization is achieved by the label $f_C : \mathcal{Z} \mapsto C(e_i)$. We can measure $\mathcal{L}$'s generalization accuracy by comparing the two labels, l and f_C. Let $\mathcal{D}_Z(e_i)$ be the conditional probability of sampling $e_i \in \mathcal{Z}$, given θ.[6] If $l(e_i)$ is the class assigned to e_i by l (i.e., $l(e_i)$ is the value of the i^{th} bit of the binary vector l in Equation 4 above), then:

$$\mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta) = \sum_{e_i\in\mathcal{Z}}[\mathcal{D}_Z(e_i)\cdot\delta(l(e_i), f_C(e_i))]$$
$$\text{where } \delta(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

Computing the average generalization accuracy: $\mathcal{AGA}_{\mathcal{L}}$ is calculated by summing $\mathcal{L}$'s total generalization over all targets and dividing by $|C|$, the total number of targets (see Equation 3). For k unseen cases, there are 2^k distinct ways of assigning classes to cases in $\mathcal{Z}$; each assignment (label) represents a distinct "target concept" and $|C| = 2^k$. Let $\mathcal{F}$ be the set of all possible labels of $\mathcal{Z}$, and let $f_j \in \mathcal{F}, 1 \leq j \leq 2^k$.

For clarity, we shall drop references to θ, $\mathcal{D}$, and $\mathcal{L}$ wherever possible, with the understanding that $\mathcal{Z}$, k, and $\mathcal{F}$ are functions of θ, and that $\mathcal{L}$ learns a single label, $l = \mathcal{L}(\theta)$. Therefore, for all $\mathcal{D}$, θ, and $\mathcal{L}$, the average generalization accuracy of $\mathcal{L}$ over all target concepts is defined as:

$$\begin{aligned} \mathcal{AGA}_{\mathcal{L}} &= \frac{1}{|C|}\sum_{f_j\in\mathcal{F}}\mathcal{GA}(\mathcal{L},\mathcal{D},C,\theta) \\ &= \frac{1}{2^k}\sum_{f_j\in\mathcal{F}}\sum_{e_i\in\mathcal{Z}}[\mathcal{D}_Z(e_i)\cdot\delta(l(e_i), f_j(e_i))] \; (6) \end{aligned}$$

2.3 UNDERSTANDING THE CONSERVATION LAW

In this subsection, we show that $\mathcal{AGA}_{\mathcal{L}}$ is always 0.5. Interchanging the order of summation in Equation 6 and rearranging the terms, we get:

$$\mathcal{AGA}_{\mathcal{L}} = \frac{1}{2^k}\sum_{e_i\in\mathcal{Z}}\sum_{f_j\in\mathcal{F}}[\mathcal{D}_Z(e_i)\cdot\delta(l(e_i), f_j(e_i))]$$
$$= \sum_{e_i\in\mathcal{Z}}(\mathcal{D}_Z(e_i)\cdot\left[\frac{1}{2^k}\sum_{f_j\in\mathcal{F}}\delta(l(e_i), f_j(e_i))\right])$$

The term within the square brackets [] above averages the accuracy of l on a single case, e_i, over all 2^k labels (targets) in $\mathcal{F}$; therefore, that term can be thought of as the average generalization accuracy of $\mathcal{L}$ (or l) on that case.

$$\begin{aligned} \mathcal{AGA}_{\mathcal{L}} &= \sum_{e_i\in\mathcal{Z}}[\mathcal{D}_Z(e_i)\cdot\mathcal{AGA}_{\mathcal{L}}(e_i)] & (7) \\ \mathcal{AGA}_{\mathcal{L}}(e_i) &= \frac{1}{2^k}\sum_{f_j\in\mathcal{F}}\delta(l(e_i), f_j(e_i)) & (8) \end{aligned}$$

[4] As θ can contain repeats in its n samples, $k \geq (m-n)$.

[5] Whereas l typically maps all m cases in $(\mathcal{Z}\bigcup\theta)$ to $\{C_0, C_1\}$, recall that for generalization purposes we are only interested in l's mapping from $e_i \in \mathcal{Z}$.

[6] If the sampling distribution for "testing" is iid and is the same as the "training" distribution, then $\mathcal{D}_Z$ may be derived from $\mathcal{D}$ by uniformly normalizing $\mathcal{D}$ for all $e_i \in \mathcal{Z}$. However, for many scenarios, such as active learning, $\mathcal{D}_Z$ may be very different from $\mathcal{D}$.

where $\mathcal{AGA}_{\mathcal{L}}(e_i)$ is the average generalization accuracy of $l(=\mathcal{L}(\theta))$ on a single case in $\mathcal{Z}$ (averaged over all target concepts). Equation 7 is an intuitive restatement of the generalization accuracy of a learner as being the weighted sum of the generalization accuracy of the learner on individual cases, weighted by the probability of drawing each case in $\mathcal{Z}$.

Consider Equation 8 above. Assume, without loss of generality, that l labels an arbitrary unseen case, e_i, to be C_0. Notice that half the 2^k concepts $f_j \in \mathcal{F}$ classify e_i to be C_0 (for these $\delta(l(e_i), f_j(e_i)) = 1$) and the other half classify e_i to be C_1 (for these $\delta(l(e_i), f_j(e_0)) = 0$). From Equation 8, $\mathcal{AGA}_{\mathcal{L}}(e_i)$ reduces to 0.5 for all cases. This is also true for all possible labels, l, and for all possible assignments of cases to θ and $\mathcal{Z}$. Therefore: $\forall e_i \in \mathcal{Z}, \mathcal{AGA}_{\mathcal{L}}(e_i) = 0.5$. Recall that other than the three simplifying assumptions made in Section 2.2 (which we will relax in Appendix A), no restrictions of any kind have been placed on θ, $\mathcal{D}$, n, C, and $\mathcal{L}$. Therefore, noting that $\sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) = 1$, Equations 7 and 8 reduce to:

$$\begin{aligned} \forall \mathcal{L}, \mathcal{D}, n, \theta \textit{ of size } n, \quad \mathcal{AGA}_{\mathcal{L}}(\mathcal{D}, \theta) &= 0.5, \\ \text{and } \forall e_i \in \mathcal{Z}, \quad \mathcal{AGA}_{\mathcal{L}}(e_i) &= 0.5 \qquad (9) \end{aligned}$$

From Equations 7 and 8, the CLGP (Equation 3), can be rewritten $\forall \mathcal{L}$ as (Note $\sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) = 1$):

$$\sum_{\mathcal{D}, n, \theta} \textit{prob}(\theta) \cdot \{ \sum_{e_i \in \mathcal{Z}} [\mathcal{D}_Z(e_i) \cdot \mathcal{AGA}_{\mathcal{L}}(e_i)] - 0.5 \} = 0$$

$$\sum_{\mathcal{D}, n, \theta} \textit{prob}(\theta) \cdot \sum_{e_i \in \mathcal{Z}} [\mathcal{D}_Z(e_i) \cdot \{\mathcal{AGA}_{\mathcal{L}}(e_i) - 0.5\}] = 0 \quad (10)$$

Because $\mathcal{AGA}_{\mathcal{L}}(e_i)$ always equals 0.5 (Equation 9), every term in the innermost summation $\{\}$ is identically 0. Therefore, the following very simple statement is equivalent to the CLGP: Given one (unseen) case, labeling that case to be of class C_0 (or C_1) will result in zero generalization performance when generalization is summed uniformly over the two possible classifications $\{C_0, C_1\}$ for that case.

3 AN ALTERNATE MEASURE OF GENERALIZATION

Every labeling of an unseen case does give a generalization performance of 0 when summed over both possible classes. While this statement may not be particularly interesting, it certainly is true; what does interest us is the impact, if any, of this statement on machine learning. Schaffer [1994] states that "Roughly speaking, the [CLGP] result indicates that generalization is a zero-sum enterprise - for every performance gain in some subclass of learning situations there is an equal and opposite effect in others," However, we find ourselves in the position of believing that induction is certainly far from hopeless in our universe, and also admitting that the CLGP is (trivially) true. Therefore, it must be the case that the CLGP, while true, is not really applicable to "learning" as we are interested in it.

Suppose we consider more carefully the analogy of the CLGP being similar to physical laws of conservation, such as the conservation of momentum and energy and the "equal and opposite reaction" force law. These laws do not consider the distribution over the units being conserved and, as with the CLGP, simply count the units. This is reasonable for physical laws because (for the purposes of any particular conservation law) these units are considered to be indistinguishable from each other; that is, one unit of momentum is equivalent to and indistinguishable from any other unit of momentum (as any joule is equivalent to any other joule). However, there is no reason to expect that all concepts are in any sense equivalent or indistinguishable. "Pure" induction is impossible – if one does not distinguish between concepts. This statement, however, has little meaning in the real world, where the distribution over target concepts is an important piece of information (that is ignored by the CLGP).

For example, when Schaffer [1994] shows "how the conservation law applies to a real learner," i.e., the majority learner, he ignores the distribution over target concepts. Once these distributions are considered, however, it is not hard to create learners with net positive (or negative) generalization. In this section, our goal is to include information about distributions over target concepts and examine the Conservation Law in light of this information. We do this by presenting a different metric, $\mathcal{EGA}$, that measures a learner's *expected generalization accuracy.* We compare $\mathcal{EGA}$ with the CLGP's (implicit) $\mathcal{AGA}$ measure, which sums (averages) generalization accuracy over all target concepts. We then pose the query: "Under what conditions will expected generalization performance be conserved?" The answer allows us to determine the class of universes in which learning is impossible.

3.1 EXPECTED GENERALIZATION

The real world is reflected by a target concept, C, that (correctly) classifies all m cases in the problem domain for any particular experiment. Specifically, we are interested in the target label, $f_C : \mathcal{Z} \mapsto C(e_i)$. What we truly want to measure is not how well a learner can generalize over all possible concepts, but how well the learner actually does with respect to the actual target, f_C. Recall from Equation 5, that for an arbitrary θ and $\mathcal{D}$, the generalization accuracy of $\mathcal{L}$ is:

$$\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) = \sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) \cdot \delta(l(e_i), f_C(e_i))$$

The problem, obviously, is that we do not know f_C; in fact, each one of the labels $f_j \in \mathcal{F}$ is a potential candidate for f_C. (Recall from Section 2.2 that $\mathcal{F}$ was

Table 1: A non-uniform distribution of target concepts, $\mathcal{P}_Z$, and the expected generalization performance ($\mathcal{EGP}$) of different learners based on their predictions on the test cases, over all target concepts. Each cell of the main part of this table contains the generalization performance of a learner (column) for a particular target concept (row). In the final column, the generalization performances for each concept, weighted by the $\mathcal{D}_Z(e_i)$'s, are summed to calculate the $\mathcal{EGP}$.

$\mathcal{L}(\theta)$	$\{\sum_{e_i} \mathcal{D}_Z(e_i)[\delta(l, f_j) - 0.5]\}$				$\sum_{f_j} \mathcal{P}_Z\{.\}$
$= l$	$\mathcal{P}_Z(++) = 0.4$	$\mathcal{P}_Z(--) = 0.4$	$\mathcal{P}_Z(+-) = 0.1$	$\mathcal{P}_Z(-+) = 0.1$	$= \mathcal{EGP}$
$++$	$+0.5$	-0.5	0	0	0
$--$	-0.5	$+0.5$	0	0	0
$+-$	0	0	$+0.5$	-0.5	0
$-+$	0	0	-0.5	$+0.5$	0

defined as the set of the 2^k possible labels of $\mathcal{Z}$.) So far the discussion on the CLGP has ignored the distribution of target concepts. Let $\mathcal{P}$ denote the distribution of target concepts over all m possible cases; given a specific θ, we are interested in $\mathcal{P}_Z$, the distribution of target concepts over the corresponding $\mathcal{Z}$ conditioned on θ, the information that we have already seen.[7] Let $\mathcal{P}_Z(f_j)$ be the conditional probability that a particular label f_j is the target label, conditioned on θ. Since we have no way of knowing for certain which f_j is the target label (for the current experiment), the best we can measure is the expected value of the generalization accuracy of a learner: the *expected generalization accuracy*, $\mathcal{EGA}_\mathcal{L}$.

$$\mathcal{EGA}_\mathcal{L} = \sum_{f_j \in \mathcal{F}} (\mathcal{P}_Z(f_j) \cdot \sum_{e_i \in \mathcal{Z}} [\mathcal{D}_Z(e_i) \cdot \delta(l(e_i), f_j(e_i))]) \qquad (11)$$

Before we examine this new measure, let us compare $\mathcal{EGA}_\mathcal{L}$ with the CLGP's average generalization accuracy metric, $\mathcal{AGA}_\mathcal{L}$. From Equation 6:

$$\mathcal{AGA}_\mathcal{L} = \sum_{f_j \in \mathcal{F}} (\frac{1}{2^k} \cdot \sum_{e_i \in \mathcal{Z}} [\mathcal{D}_Z(e_i) \cdot \delta(l(e_i), f_j(e_i))]) \quad (12)$$

Comparing Equation 11 with Equation 12 we notice that the CLGP implicitly assumes that $\mathcal{P}_Z(f_j) = 1/2^k$ for all possible targets, $f_j \in \mathcal{F}$. This corresponds to the uniform concept distribution, $\mathcal{P}_Z$*-random*, in which every possible classification of unseen cases is equally likely. This is the definition of a uniformly random universe, in which learning is impossible.[8]

3.2 WHEN IS EXPECTED GENERALIZATION CONSERVED?

The pertinent question for real-world data is: When is the *expected* generalization performance equal to 0? That is, *under what conditions will all learners have zero $\mathcal{EGP}$?* Recall Equation 2 from Section 2 that restates the CLGP:

$$\forall \mathcal{L} \sum_{\mathcal{D}} \sum_{n} \sum_{\theta} (prob(\theta) \cdot \sum_{f_j \in \mathcal{F}} [\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5]) = 0$$

A condition for conserving $\mathcal{EGP}(= \mathcal{EGA} - 0.5)$ would have the form:[9]

$$\forall \mathcal{L} \sum_{\mathcal{D}} \sum_{n} \sum_{\theta} (prob(\theta) \cdot [\mathcal{EGA}_\mathcal{L} - 0.5]) = 0 \qquad (13)$$

The degenerate situation is when $\mathcal{EGA}_\mathcal{L} = \mathcal{AGA}_\mathcal{L}$, i.e., for every $\mathcal{L}$, $\mathcal{D}$, n, θ (and e_i), $\mathcal{EGA}_\mathcal{L}$ is always 0.5 (i.e., $\mathcal{P}_Z$*-random*). There are, however, many other $\mathcal{P}_Z$ distributions for which expected generalization performance ($\mathcal{EGP}$) is conserved. All these distributions appear to share a common symmetry property. Before characterizing the conditions under which $\mathcal{EGP}$ is conserved, we first provide an example that illustrates how a non-uniform $\mathcal{P}_Z$ distribution can conserve $\mathcal{EGP}$. We abbreviate C_1 with "+" and C_0 with "−". Suppose there are two unseen cases, $\mathcal{Z} = \{e_1, e_2\}$, where

[7] Under a variety of assumptions, $\mathcal{P}_Z$ may be derived by uniformly normalizing the probabilities of all concepts in $\mathcal{P}$ that are consistent with θ. However, this, by no means, is always the case. For instance, assume that a benevolent teacher, who is aware of the target concept and has access to all the cases, provides the n samples in θ. Furthermore, assume that the teacher provides only those samples which would be classified as positive by the target concept (unless all possible positive examples have already been provided). Then the presence of a single negative example in a particular θ means that the only concept (in $\mathcal{P}_Z$) with non-zero probability of being the target is the one which labels all the unseen cases negative. Thus, $\mathcal{P}_Z$ can be very different from $\mathcal{P}$.

[8] It is the uniformly random universe that is implicitly used in the majority learner example in Schaffer [1994].

[9] Just as we weight θ with $prob(\theta)$, the probability of drawing the n samples in the training set for a given n and $\mathcal{D}$, we could also weight Equation 13 with $prob(n)$, the probability that we will choose to draw n samples from a given $\mathcal{D}$, and with $prob(\mathcal{D})$, the probability that $\mathcal{D}$ is the sampling distribution. For the purposes of this paper we will assume that $prob(n)$ and $prob(\mathcal{D})$ are uniform. Future work will relax this assumption.

$\mathcal{D}_Z(e_1) = \mathcal{D}_Z(e_2) = 0.5$. The second row of Table 1 displays $\mathcal{P}_Z$, the probability distribution for the four possible target concepts over the unseen cases. Note that this is *not* a uniform $\mathcal{P}_Z$ distribution; however each unseen case is equally likely to be positive or negative. For example, the probability that e_1 is positive is equal to $\mathcal{P}_Z(++) + \mathcal{P}_Z(+-) = 0.5$, and the probability that e_1 is negative is equal to to $\mathcal{P}_Z(-+) + \mathcal{P}_Z(--) = 0.5$.

Consider four possible learners which produce the labels $\langle ++ \rangle$, $\langle -- \rangle$, $\langle +- \rangle$, $\langle -+ \rangle$, respectively.[10] Each learner corresponds to a row of Table 1, and each target concept corresponds to a column. Recall from Equation 5 that a learner gets an accuracy score of $+1$ for each correct prediction, and 0 for each incorrect prediction. This corresponds to a generalization performance score $\{\delta(l(e_i), f_j(e_i)) - 0.5\}$ of $+0.5$ and -0.5 for a correct and incorrect prediction, respectively. The bottom 4 rows of Table 1 display the generalization performance for each of the four learners with respect to each of the four target concepts; that is, the above performance score weighted by $\mathcal{D}_Z(e_i)$ for each (l, f_j) pair. The final column in the Table 1 presents the $\mathcal{EGP}$ ($= \mathcal{EGA}_{\mathcal{L}} - 0.5$): the sum of the different generalization performances weighted by $\mathcal{P}_Z(f_j)$. Notice that $\mathcal{EGP}$ is zero for every learner; thus, generalization is impossible for this $\mathcal{P}_Z$.[11]

Substituting the value of $\mathcal{EGA}_{\mathcal{L}}$ from Equation 11 into Equation 13 and rearranging, the condition for conservation of $\mathcal{EGP}$ can be rewritten as:

$$\forall \mathcal{L} \sum_{\mathcal{D}} \sum_{n} \sum_{\theta} (prob(\theta) \cdot \sum_{f_j \in \mathcal{F}} (\mathcal{P}_Z(f_j) \cdot \sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) \cdot [\delta(l(e_i), f_j(e_i)) - 0.5]) = 0$$

$$\forall \mathcal{L} \sum_{\mathcal{D}} \sum_{n} \sum_{\theta} (prob(\theta) \cdot \sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) \cdot \sum_{f_j \in \mathcal{F}} (\mathcal{P}_Z(f_j) \cdot [\delta(l(e_i), f_j(e_i)) - 0.5]) = 0 \quad (14)$$

The question, of course, is when this holds. Although the answer to this question is mathematically obvious, in the remainder of this section we elaborate situations for which $\mathcal{EGP}$ is conserved (Equation 14) on a case-by-case basis. We do so in order to elucidate the types of symmetries that would result in zero $\mathcal{EGP}$. Our motivation for this elaboration is that associated with each situation is an open research question as to whether this type of generalization performance symmetry holds in our world. Obviously, one particular set of cases when $\mathcal{EGP}$ is conserved is when the innermost summation of Equation 14 is always zero.

$$\forall \mathcal{L}, \forall \mathcal{D}, \forall n, \forall \theta, \forall e_i \in \mathcal{Z}(\theta) \qquad \sum_{f_j \in \mathcal{F}} \{\mathcal{P}_Z(f_j) \cdot [\delta(l(e_i), f_j(e_i)) - 0.5]\} = 0$$

There are a number of ways to satisfy these equations. The most obvious, mentioned above, is when each unseen case is equally likely to be positive or negative. Let $\mathcal{Q}(l, e_i) = \sum_{f_j} (\mathcal{P}_Z(f_j)(\delta(l(e_i), f_j(e_i)) - 0.5))$, or more simply $\mathcal{Q}$, represent the expected generalization performance of a learned label l on a single unseen case e_i. Then examining Equation 14, another way to achieve zero expected generalization performance is when $\sum(\mathcal{D}_Z \cdot \mathcal{Q}) = 0$. Therefore, if it is not the case that both classifications of test cases are equally likely, it is still possible for $\mathcal{EGP}$ to be conserved if the following symmetry property exists: For some partition of the unseen cases into two subgroups, $\mathcal{Z}_1$ and $\mathcal{Z}_2$, $\sum_{e_i \in \mathcal{Z}_1} [\mathcal{D}_Z(e_i) \cdot \mathcal{Q}] = -\sum_{e_i \in \mathcal{Z}_2} [\mathcal{D}_Z(e_i) \cdot \mathcal{Q}]$. (Notice that if this is true for one partition, it is true for *every* partition.)

We next consider cases where the outer three summations $\sum_D \sum_n \sum_\theta$ contribute toward satisfying Equation 14. In other words, the inner sums of Equation 14 might be nonzero, but one or more of the three outer summations might produce a final result of 0. We consider the outer sums of Equation 14 one by one, beginning with the innermost one $\sum_\theta$. For $\mathcal{EGP}$ to be conserved for all $\mathcal{L}$, $\mathcal{D}$, and n, it must be the case that for some (every) partition of all possible training sets into two classes, Θ_1 and Θ_2, the following must hold:

$$\sum_{\theta \in \Theta_1} prob(\theta) \sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) \cdot \mathcal{Q} = -\sum_{\theta \in \Theta_2} prob(\theta) \sum_{e_i \in \mathcal{Z}} \mathcal{D}_Z(e_i) \cdot \mathcal{Q}$$

In other words, if we consider all possible training sets of size n, positive expected generalization performance on some unseen (test) sets needs to be exactly counterbalanced by negative expected generalization performance on the remaining possible unseen (test) sets (when weighted by the conditional probabilities of the θs). Likewise, for some (every) partition of training set sizes n into N_1 and N_2, $\sum_{n \in N_1}[\cdots] = -\sum_{n \in N_2}[\cdots]$ must hold for the $\mathcal{EGP}$ to be conserved if the inner sums (including $\sum_\theta$) are nonzero. Finally, for the outermost summation, $\sum_{\mathcal{D}}$, we are again faced with the

[10] This example considers only Boolean learners and Boolean targets. However, all possible learners (targets) can be expressed as weighted combinations of these learners (targets). So if $\mathcal{EGP} = 0$ for these learners, it follows that $\mathcal{EGP}$ will be conserved for *all* learners.

[11] Note that the CLGP's total generalization performance measure would simply sum each row in Table 1 without considering the weight of each target concept; this would result in zero $\mathcal{GP}$. If the probability distribution over targets, $\mathcal{P}_Z$, in Table 1 was changed, say $\mathcal{P}_Z(f_j = ++) = 1$, and $\mathcal{P}_Z(f_j) = 0$ for all other target concepts, the total $\mathcal{GP}$ would still be zero, However, the $\mathcal{EGP}$ would now be nonzero: for instance, the learner "$\langle ++ \rangle$" would have positive generalization.

same situation: for some (every) partition of the distributions into $\mathcal{D}_1$ and $\mathcal{D}_2$, generalization performance must be balanced between the two subgroups.

In summary, for each level of summation, if symmetry does not hold at some inner level but it holds at the next outer level of summation, the $\mathcal{EGP}$ will be conserved. Of course, the symmetry need not hold *within* the set of all possible θ's, n's, or $\mathcal{D}$'s (via partitioning as we just did). Alternatively, there could be a correlation between θ, n, and $\mathcal{D}$ (or some subset thereof) with respect to generalization performance. To capture this, we consider the most general conditions under which $\mathcal{EGP}$ is conserved. Earlier, we stated that a learning situation, S, is defined by the triple ($\mathcal{D}$, C, n). We now define a more specific learning situation, S^*, as $(\mathcal{D}, \mathcal{D}_Z, n, \theta, \mathcal{P}, \mathcal{P}_Z)$. Then the most general condition for the $\mathcal{EGP}$ to be conserved is: for some (every) partition of S^* into S_1^* and S_2^* the following symmetry must hold (for all learners, $\mathcal{L}$):

$$\sum_{S_1^*}[prob(\theta) \cdot \mathcal{D}_Z(e) \cdot [\mathcal{P}_Z(f) \cdot [\delta(l(e), f(e)) - 0.5]]] =$$

$$-\sum_{S_2^*}[prob(\theta) \cdot \mathcal{D}_Z(e) \cdot [\mathcal{P}_Z(f) \cdot [\delta(l(e), f(e)) - 0.5]]]$$

4 DISCUSSION

To summarize the discussion in Section 3, for $\mathcal{EGP}$ to be conserved, positive $\mathcal{EGP}$ on some subset of all possible situations must be exactly counterbalanced by negative $\mathcal{EGP}$ on the remaining possible situations. This is the same argument that has been presented in [Schaffer, 1994] regarding $\mathcal{GP}$; here, however, we present the argument in terms of a real-world measure, namely, the $\mathcal{EGP}$. We have demonstrated that $\mathcal{EGP}$, unlike the CLGP, is not trivially conserved but instead raises an intriguing and important question: What about the existence of the symmetries required for zero-$\mathcal{EGP}$ in our universe? Although the CLGP cannot be empirically tested in the real world – because it ignores the practical consideration of distributions – the $\mathcal{EGP}$ *can* be. From the evidence to date, there is no reason to believe that $\mathcal{P}_Z$ (or $\mathcal{P}$) has precise symmetry properties. (Furthermore, it is virtually tautological that learning is possible in our universe.) We rarely find test sets for which it is random whether an instance will be positive or negative. Nor do we typically find learners whose good performance on some real-world test sets is exactly counterbalanced by equally poor performance on others. Nevertheless, a very interesting open question is how *close* our universe comes to matching the required symmetry for zero $\mathcal{EGP}$, where we propose:

$$\mathcal{TEGP} =$$

$$\sum_{\mathcal{D},n,\theta}(prob(\theta) \cdot \sum_{f_j \in \mathcal{F}}[\mathcal{P}_Z(f_j) \cdot [\mathcal{GA}(\mathcal{L}, \mathcal{D}, C, \theta) - 0.5]])$$

from Equation 14 as a measure of $\mathcal{TEGP}$, the total $\mathcal{EGP}$ of any learner, $\mathcal{L}$.

Perhaps there are natural constraints that enforce approximations of some of these symmetries and thereby make the $\mathcal{TEGP}$-measure close to 0 for every (or almost every) learner $\mathcal{L}$. If so, then learning could be considered "very difficult" in our universe and it would be hard to find one learner that predicts better than another. Alternatively, perhaps there exist natural constraints that preclude our universe from being close to having these symmetries for all learners, thus making learning a very valuable enterprise. Some physicists' theories about our world have been highly predictive of unseen data. This leads us to conjecture that our world has strong regularities, rather than being nearly random. However, only time and further testing of physical theories can refine our understanding of the nature of our universe. Scientific progress might lead to a reasonable estimate of $\mathcal{P}_Z$ in our world, for example. Until that time, we agree with Wolpert [1994] that researchers should be careful not to say Algorithm A is better than Algorithm B without mentioning that this holds with respect to a particular problem distribution.

While we do not agree with the CLGP's claim that all learners must be zero-sum generalizers, we do agree with some of the points raised in Schaffer [1994]: that the study of bias is critical, that careful studies of learners (and biases) often reveal weaknesses on some data sets as well as strengths on other data sets (e.g., see [Fisher and Schlimmer, 1988; Holte, 1993; Ade *et al.*, 1995; Brodley, 1995; Provost and Buchanan, 1995]), that it is imperative to consider priors (see [Dietterich, 1989; Buntine, 1991; Buntine, 1993; Haussler *et al.*, 1994]), and that we should focus on off-training set error (see [Wolpert, 1992; Wolpert, 1994]).

In addition to the focus on off-training set error, Wolpert [1994] also points out other ways in which his results differ from prior learning theory, in particular, PAC [Valiant, 1984]. First, PAC analyses typically assume the concept class is known a priori; the CLGP does not. Second, PAC is concerned with whether it takes polynomial or exponential time to achieve a desired predictive accuracy; the CLGP boils down to a question of predictive accuracy on a single unseen instance.

In a nutshell, the underlying assumptions of the CLGP can be summarized as stating that *all of our observations in the past have no bearing/relation to what we will see in the future.* This is patently *not* a PAC assumption.

In this paper we have shown that the CLGP is equivalent to the statement: classifying an unseen case as positive (or negative) results in an generalization accuracy of 0.5, when the accuracy is measured uniformly against both possible classifications (i.e., positive or negative). Once you sum generalization uniformly over all concepts, the total generalization is always zero in *every* learning situation. Therefore, it is redundant to sum over all learning situations (i.e., over all distributions and training set sizes as done by the CLGP in Equation 1). We have also introduced the notion of expected generalization performance, or $\mathcal{EGP}$, and presented a corresponding measure which allows us to compute the $\mathcal{EGP}$ for any learner. This measure is zero for all learners, if and only if our universe is symmetric as defined in Section 3. If this symmetry does not exist, non-zero $\mathcal{EGP}$ is possible and one learner can be better than another.

Acknowledgments

We gratefully acknowledge receiving helpful comments from Thomas Hancock, Ralph Hartley, Steven Salzberg, Geoffrey Towell, and the anonymous reviewers. We also thank Michael Pazzani for starting a discussion on the Conservation Law on the Machine Learning List which he maintains.

References

[Ade *et al.*, 1995] Ade, H.; Raedt, L. De; and Bruynooghe, M. 1995. Declarative bias for specific-to-general ilp systems. *Machine Learning.* (To appear).

[Brodley, 1995] Brodley, C. 1995. Recursive automatic bias selection for classifier construction. *Machine Learning.* (To appear).

[Buntine, 1991] Buntine, W. 1991. Theory refinement of Bayesian networks. In D'Ambrosio, B.; Snets, P.; and Bonissone, P., editors 1991, *Uncertainity in Artificial Intelligence: Proceedings of the Seventh Conference.*

[Buntine, 1993] Buntine, W. 1993. Prior probabilities: A tutorial and unifying view. Technical report, RIACS/NASA Ames Research Center.

[Dietterich, 1989] Dietterich, T. 1989. Limitations on inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning.* 124–128.

[Fisher and Schlimmer, 1988] Fisher, D. and Schlimmer, J. 1988. Concept simplification and prediction accuracy. In *Proceedings of the Fifth International Conference on Machine Learning.* 22–28.

[Haussler *et al.*, 1994] Haussler, D.; Kearns, M.; and Schapire, R.E. 1994. Bounds on the sample complexity of Bayesian learning using informations theory and the VC dimension. *Machine Learning* 14:83–113.

[Holte, 1993] Holte, R. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11(1):63–90.

[Mitchell, 1980] Mitchell, T.M. 1980. The need for biases in learning generalizations. Technical Report CBM-TR-117, Computer Science Department, Rutgers University, New Brunswick, NJ.

[Provost and Buchanan, 1995] Provost, F. and Buchanan, B. 1995. Inductive policy: The pragmatics of bias selection. *Machine Learning.* (To appear).

[Rao *et al.*, 1995] Rao, R.B.; Gordon, D.; and Spears, W. 1995. On the conservation of generalization and expected generalization. Technical Report AIC-95-006, Naval Research Laboratory, S.W. Washington DC. (In progress).

[Schaffer, 1993] Schaffer, C. 1993. Overfitting avoidance as bias. *Machine Learning* 10:153–178.

[Schaffer, 1994] Schaffer, C. 1994. A conservation law for generalization performance. In *Proceedings of the Eleventh International Conference on Machine Learning.* 259–265.

[Utgoff, 1986] Utgoff, P. E. 1986. Shift of bias of inductive concept learning. In Michalski, R.S.; Carbonell, J.G.; and Mitchell, T.M., editors 1986, *Machine Learning: An Artificial Intelligence Approach, Vol II.* Morgan Kaufmann Publishers. 107–148.

[Valiant, 1984] Valiant, L.G. 1984. A theory of the learnable. *Communications of the ACM* 27:1134–1142.

[Wolpert, 1992] Wolpert, D.H. 1992. On the connection between in-sample testing and generalization error. *Complex Systems* 6:47–94.

[Wolpert, 1994] Wolpert, D. H. 1994. Off-training set error and a priori distinctions between learning algorithms. Technical report, Santa Fe Institute, Santa Fe, NM.

A RELAXING THE SIMPLIFYING ASSUMPTIONS

Here, we sketch rough proofs showing that our restatement of the CLGP holds even after we relax the simplifying assumptions we made in Section 2.1. In other words, zero-sum $\mathcal{GP}$ depends only on summing up generalization over all target concepts, and is independent of the earlier assumptions. We will now permit $\mathcal{L}$ to learn multiple labels, and we will allow these labels and the target concepts to be stochastic.

Let a learner probabilistically learn labels, l_j, each with some probability, w_j ($\sum w_j = 1$). From Equation 9, $\forall l_j$, $\mathcal{AGA}_{l_j} = 0.5$; therefore, the weighted sum of all these $\mathcal{AGA}_{l_j}$'s is also 0.5 ($\mathcal{AGA}_{\mathcal{L}} = \sum w_j \mathcal{AGA}_{l_j} = 0.5 \sum w_j = 0.5$).

Now, let $\mathcal{L}$ learn a stochastic label, l: we expand the (Boolean) definition of a label in Equation 4 to be a vector of k real numbers, $b_i \in [0,1]$, and view each b_i as the probability that l classifies e_i as C_1 (earlier $b_i \in \{0,1\}$ in Equation 4). Similarly, we change the definition of $\delta(a,b)$ so that $\delta(a,b) = 1 - |a-b|$ (for $a, b \in \{0,1\}$, this reduces to Equation 5). Then summing over the two labels $\{C_0, C_1\}$ for each e_i gives 0.5 for any value of b_i: $\mathcal{AGA}_{\mathcal{L}}(e_i) = (\delta(b_i,0) + \delta(b_i,1))/2 = (1 - b_i + 1 - (1 - b_i))/2 = 0.5$. (Alternately, we can represent a stochastic label as multiple Boolean labels, l_j, with varying weights, w_j, ($\sum w_j = 1$) which also results in an $\mathcal{AGA}_{\mathcal{L}} = 0.5$ as discussed earlier.)

Finally consider the situation where the target concept, C, itself is stochastic. Assume that each $e_i \in \mathcal{Z}$ is assigned to C_1 with probability q_i and to C_0 with $(1 - q_i)$. Summing over all concepts is interpreted by the CLGP as meaning that q_i varies *uniformly* over [0,1]. Then by integrating over the interval, [0,1], we get $\forall e_i, \mathcal{AGA}_{\mathcal{L}}(e_i) = 0.5$. (If $\mathcal{L}$ assigns e_i to be C_0, we get $\mathcal{AGA}_{\mathcal{L}}(e_i) = \int_0^1 \delta(0, q_i) dq_i = \int_0^1 (1 - q_i) dq_i = 0.5$; if $\mathcal{L}(e_i) = C_1$, then $\mathcal{AGA}_{\mathcal{L}}(e_i) = \int_0^1 \delta(1, q_i) dq_i = \int_0^1 (q_i) dq_i = 0.5$.) (Similar proofs exist in [Wolpert, 1992].) For more details, including extending the proofs to concepts that are combinations of the above situations, see [Rao *et al.*, 1995].